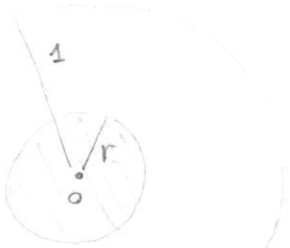
$p = 2$: $\quad x \sim \mathcal{U}[S^2(1)]$.

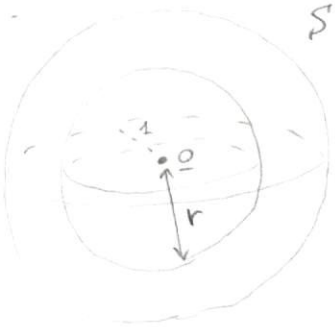
$$0.1 = \frac{\text{Area}(S^2(r))}{\text{Area}(S^2(1))} = \frac{\pi r^2}{2\pi} = r^2.$$

$$\Rightarrow r = \sqrt{\frac{1}{10}} = 0.31.$$



$S^2(1)$

$p = 3$:



$S^3(1)$

$$0.1 = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi} = r^3.$$

$$\Rightarrow r = \left(\frac{1}{10}\right)^{1/3} = 0.46.$$

If you want to capture 10 % of your friends, you have to travel 50 % of the radius of the universe.

Going on, $\begin{cases} p = 100 \Rightarrow r_{0.1} = 0.97 \\ p = 1000 \Rightarrow r_{0.1} = 0.99. \end{cases}$

The equation is: $\boxed{r^p = 0.1.}$

When p is big, we will have many isolated cases, very different one from each other.

How do we avoid the curse of dimensionality?

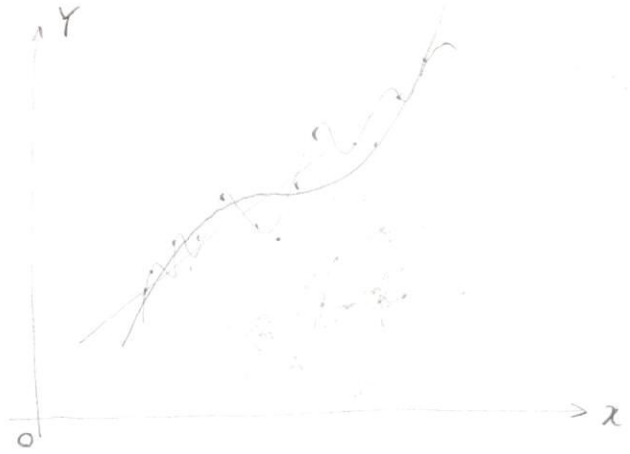
1) Reduce the dimensionality p. We could do it with Principal Component Analysis.

2) Use a parametric model:

For example, $f(\underline{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.

Everything is reduced to estimating a given small number p of parameters $(\beta_0, \dots, \beta_p)$.

The first approach is data driven.



$$Y = f(X) + \varepsilon .$$

Options:

$$f(x) = \beta_0 + \beta_1 x .$$

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 .$$

$$f(x) = \beta_0 + \beta_1 x + \dots + \beta_K x^K .$$

We have many different models.

The third one reduces completely the error, i.e. fits perfectly the data. But we see that we have a <u>variance - Bias tradeoff</u>.

$$E_{|\mathcal{X}}\left[(Y_0 - \hat{f}(x_0))^2 \right] = \left(f(x_0) - \hat{f}(x_0)\right)^2 + \mathrm{Var}(\varepsilon_0).$$

I want to take the expected value all over the set of datas $\{\mathcal{X}\}$:

$$E\left[E_{|\mathcal{X}}\left[(Y_0 - \hat{f}(x_0))^2 \right] \right] = E\left[\left(f(x_0) - \hat{f}(x_0)\right)^2 \right] + \mathrm{Var}(\varepsilon_0) =$$

($\mathcal{X}$: random itself)

$$= E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right] + E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2 \right] + \mathrm{Var}(\varepsilon_0) =$$

$$= E\left[\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2 \right] + E\left[\left(E\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2 \right] + \underbrace{2E[\dots]}_{=0} + \mathrm{Var}(\varepsilon_0) =$$

$$= \left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2 + \mathrm{Var}\left(\hat{f}(x_0)\right) + \mathrm{Var}(\varepsilon_0).$$

squared bias: $\left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2$

variability of the model: $\mathrm{Var}\left(\hat{f}(x_0)\right)$

Bias: difference from what you predict and what you observe (on average)

$$\Rightarrow \quad \boxed{E\left[(Y_0 - \hat{f}(x_0))^2 \right] = \left(f(x_0) - E\left[\hat{f}(x_0)\right]\right)^2 + \mathrm{Var}\left(\hat{f}(x_0)\right) + \mathrm{Var}(\varepsilon_0).}$$

Lab Session in R:

12/3/20

13/3/20

Presentation of Datasets:

Data requirements:

- multivariate data;
- two or more groups;
- Websites: Kaggle
  Comune di Milano
  Regione Lombardia
  ISTAT
  EUROSTAT
  NASA

Project steps:

1) Identification of stakeholders (who is interested in your analysis)
2) Identification of research questions
3) Building the data set from raw data
4) Data analysis
5) Answer to the research questions

Have in mind somebody asking you some questions (the stakeholders).

At the end, we will present a poster.

Some Possible Datasets:

(Riccardo Pelo)

1) West Nile Virus Prediction.
2) Predicting Molecular Properties.

(Agostino Tatì)

- 4) Atlas of Economic Complexity (Harvard University)
3) Corona Virus dataset.

Question: predict the spread of Corona Virus in next months.

(Alex Didkovskij)

5) Estimate the unit sales of Walmart retail goods.

6) Factors that help measure how young children learn.

(Matteo Fontana)

7) Radar Interferometry

(Mara Bernardi)

8) Spectroscopic Data for Planetary Investigations.

(Alessandra Menafoglio)

9) Earthquake intensity measures.

(Marta Pittaficco)

10) Osteosarcoma

(Daniele Faiacano)

11) Chest Radiography

- 12) Data analysis for Mobility (Agostino Tonti) (Marta Galvani)
  - ↳ Eaadtra
  - ↳ Train delays
  - ↳ changing network infrastructures

(Simone Vantini)

- 13) Safari Njema ⇒ "Buon Viaggio".

Transport Poverty: lack of mobility services

↳ Problem: Informal Mobility.

Solution: 70% of population own a mobile phone.

Maputo, Mozambico.

14) Identity Recognition via Active Shoes.

↳ "Individual Mobility" problem

16/3/20

Sample geometry:

- n : statistical units
- p : features (or variables)

$$\mathbb{X} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & & \cdots & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ & & & \\ & & & \\ x_{n1} & & \cdots & x_{np} \end{bmatrix} \end{array}$$

data matrix (data frame)

Two uses: by rows and by columns.

- $\underline{x}_i' = (x_{i1} \cdots x_{ip}) \in \mathbb{R}^p$ → i-th multivariate observation, $i = 1, \dots, n$.

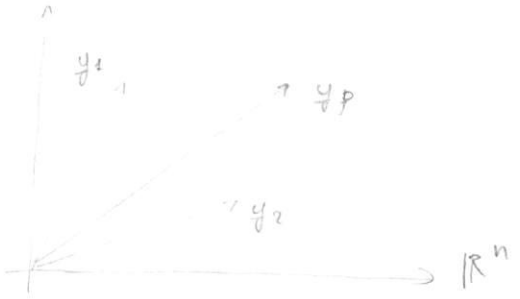
$\underline{y}_1 = (x_{11}, x_{21}, \dots, x_{n1})' \in \mathbb{R}^n$

↳ sample from X_1

⋮

- $\underline{y}_j = (x_{1j}, \dots, x_{nj})' \in \mathbb{R}^n$, $j = 1, \dots, p$

↳ sample from X_j



- We can take the mean of these samples:

$\bar{y}_j = (x_{1j}, \dots, x_{nj})' \in \mathbb{R}^n$

$$\bar{x}_j := \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \quad j = 1, \dots, p$$

→ $$\underline{\bar{x}} := \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$ sample mean vector

- Variance:

$$s_{jj} := \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$ sample variance of $\underline{x}_j$

The standard deviation of x_j is $\sqrt{s_{jj}}$.

- Covariance between X_k, X_j:

$$s_{kj} := \mathrm{Cov}(\underline{x}_k, \underline{x}_j) = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j). \quad k, j = 1, \dots, p.$$

So that $\mathrm{Cov}(\underline{x}_j, \underline{x}_j) = s_{jj}$.

So we can introduce the covariance matrix (sample):

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & s_{pp} \end{pmatrix}$$

Sample variance/covariance matrix

Properties of S:

i) symmetric.

- Correlation: $$r_{kj} := \frac{s_{kj}}{\sqrt{s_{kk}\, s_{jj}}} = Cor(X_k, X_j), \quad k, j = 1, \dots, p.$$

In can collect all the informations in a matrix:

$$r := \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ & \ddots & \vdots \\ & & r_{pp} \end{pmatrix}$$

Sample correlation matrix

We know that $r_{kj} \in [-1, 1]$.

What do we do with mean and standard deviation?

Example: X_1 : height of an Italian male.

$\bar{x}_1 = 1.80$ m, $\sqrt{s_{11}} = 0.03$ m.

We know that at least 90% of the population is inside the interval:

$$[\bar{x}_1 - 3\sqrt{s_{11}},\ \bar{x}_1 + 3\sqrt{s_{11}}] = [1.74, 1.86].$$

This fact comes from Chebyshev inequality:

$$Fr\left[\bar{x}_1 - k\sqrt{s_{11}} \le X_1 \le \bar{x}_1 + k\sqrt{s_{11}}\right] \ge 1 - \frac{1}{k^2}.$$

$k = 3$: $Fr\left[\bar{x}_1 - 3\sqrt{s_{11}} \le x_1 \le \bar{x}_1 + 3\sqrt{s_{11}}\right] \ge \frac{8}{9}$.

$k = 2$: $Fr[\cdots] \ge \frac{3}{4}$.

This without any hypothesis on the distribution of X_1.

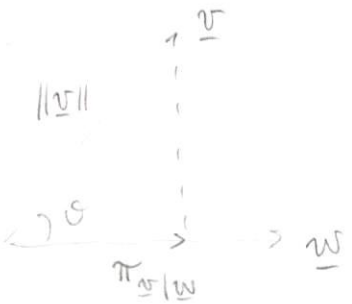
A bit of geometry:

$\underline{v}, \underline{w} \in \mathbb{R}^n$. In $\mathbb{R}^n$ we have an inner product:

$$\boxed{\langle \underline{v}, \underline{w} \rangle := \underline{v}' \underline{w}}.$$

We also have a notion of length:

$$\boxed{\|\underline{v}\| := \sqrt{\langle \underline{v}, \underline{v} \rangle} = \sqrt{\sum v_i^2}}.$$



Moreover, we have the notion of angle:

$$\boxed{\cos\vartheta := \frac{\langle \underline{v}, \underline{w} \rangle}{\|\underline{v}\| \, \|\underline{w}\|} = \frac{\sum v_i w_i}{\sqrt{(\sum v_i^2)(\sum w_i^2)}}}.$$

Projection of $\underline{v}$ on $\underline{w}$:

$$\pi_{\underline{v}|\underline{w}} = \pi_{\underline{v}|\mathcal{L}(\underline{w})} = \|\underline{v}\| \cos\vartheta \cdot \frac{\underline{w}}{\|\underline{w}\|} =$$

$\mathcal{L}(\underline{w}) = \{\underline{z} : \underline{z} = c\underline{w}, c \in \mathbb{R}\}$.

$$= \frac{\|\underline{v}\| \, \underline{v}'\underline{w}}{\|\underline{v}\| \|\underline{w}\|} \cdot \frac{\underline{w}}{\|\underline{w}\|} = \frac{\underline{v}'\underline{w}}{\|\underline{w}\|^2} \cdot \underline{w} =$$

$$= \frac{\underline{w}'\underline{v}}{\underline{w}'\underline{w}} \cdot \underline{w} = \frac{\underline{w}\,\underline{w}'}{\underline{w}'\underline{w}} \cdot \underline{v}.$$

this is a matrix, that projects $\underline{v}$ on $\mathcal{L}(\underline{w})$, called orthogonal projection.

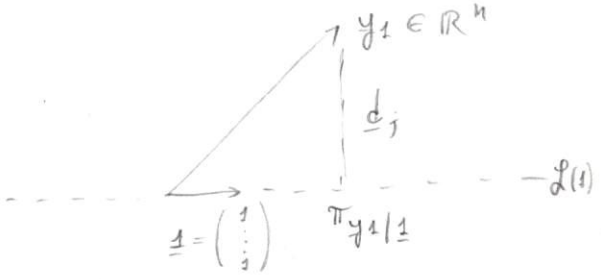
Having introduced this geometry, let's go back to our data sample.

$$\mathbb{X} = [y_1, \dots, y_p], \quad y_j \in \mathbb{R}^n \text{ sample from } X_j, \; j = 1, \dots p.$$

Take, for instance, y_1 as sample of heights.

Consider now the $\mathcal{L}(\underline{1})$ space, space of "no statistics" (space where every sample is equal):

$$\underline{v} \in \mathcal{L}(\underline{1}) \iff \underline{v} = c \cdot \underline{1}, \quad c \in \mathbb{R}.$$



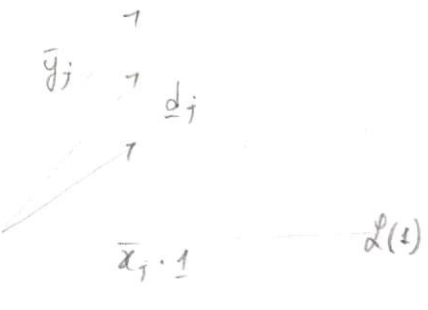
Let's project y_1 in $\mathcal{L}(\underline{1})$.

$$\pi_{y_1|\underline{1}} = \frac{\underline{1} \cdot \underline{1}'}{\underline{1}'\underline{1}} \cdot \underline{y}_1 = \frac{\sum_{i=1}^n x_{i1}}{n} \underline{1} = \bar{x}_1 \cdot \underline{1} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_1 \end{pmatrix}.$$

Hence, geometrically the mean is the projection of the vector in the space with no variability.

anyway, very different vectors can have the same mean, but will have a different deviation $\underline{d}_j$.

$$\underline{d}_j := \underline{y}_j - \bar{x}_j \cdot \underline{1} = \begin{pmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{pmatrix} \quad \text{deviation}$$



The longer is the vector $\underline{d}_j$, the greater is the norm of $\underline{d}_j$, the worse is the approximation.

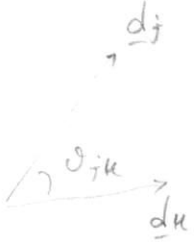
$$\|\underline{d}_j\| = \sqrt{\underline{d}_j' \underline{d}_j} = \sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

$$= \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} = \sqrt{n}\, \sqrt{s_{jj}}.$$

Take now two vectors $\underline{y}_j$, $\underline{y}_k$.

$$\begin{cases} \underline{y}_j = \bar{x}_j \underline{1} + \underline{d}_j \\ \underline{y}_k = \bar{x}_k \cdot \underline{1} + \underline{d}_k \end{cases}$$

$\bar{x}_j \underline{1}, \bar{x}_k \underline{1} \in \mathcal{L}(\underline{1})$; $\underline{d}_j, \underline{d}_k \in \mathcal{L}^{\perp}(\underline{1})$



Consider the angle between $\underline{d}_j$, $\underline{d}_k$, say ϑ_{jk}.

- $\vartheta_{jk} = 0 \Rightarrow \underline{d}_j \in \mathcal{L}(\underline{d}_k) \Rightarrow \underline{d}_j = \beta \underline{d}_k$, $\beta \in \mathbb{R}$

$$\Rightarrow \underline{y}_j - \bar{x}_j \cdot \underline{1} = \beta (\underline{y}_k - \bar{x}_k \underline{1})$$

$$\Rightarrow \underline{y}_j = \bar{x}_j \underline{1} + \beta \underline{y}_k - \beta \bar{x}_k \underline{1}.$$

Hence, we have a perfectly linear relationship between the two variables.

- $\vartheta_{jk} = \frac{\pi}{2}$: no information in $\underline{d}_k$ related to $\underline{d}_j$.

- $\vartheta_{jk} \in (0, \frac{\pi}{2})$: there is some information contained in $\underline{d}_k$, and some information that is left out.

Therefore, we need to compute the angle ϑ_{jk} in order to know how much one variable is related to the other.
More precisely, we want to measure $\cos \vartheta_{jk}$:

$$\cos \vartheta_{jk} = \frac{\underline{d}_j' \cdot \underline{d}_k}{\|\underline{d}_j\| \, \|\underline{d}_k\|} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\left(\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2\right)\left(\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2\right)}} =$$

$$= \frac{s_{jk}}{\sqrt{s_{jj}\, s_{kk}}} = \text{Corr}(X_j, X_k) = r_{jk} \Rightarrow \text{Therefore it's natural that } r_{jk} \in [-1, 1].$$

- $\vartheta_{jk} = 0 \Rightarrow \cos\vartheta_{jk} = 1 \Rightarrow r_{jk} = 1 \Rightarrow \underline{d}_j \in \mathcal{L}(\underline{d}_k)$.
- $\vartheta_{jk} = \frac{\pi}{2} \Rightarrow \dots \underline{d}_j \perp \underline{d}_k$.

Let's see now the data matrix from the other perspective:

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix}, \quad \underline{x}_i \in \mathbb{R}^p, \quad i = 1, \dots, n.$$

$\underline{x}_i$ is the realization of a random vector $\underline{X}_i$.

(Basic) General assumption: $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim \underline{X} \in \mathbb{R}^p$.

We need some tools to work with random vectors.

Random Vectors:

Let $\underline{X}$ random vector in $\mathbb{R}^p$:

$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}, \text{ with } X_k, k \in \{1, \dots, p\} \text{ random variable} \in \mathbb{R}.$$

- Probability law of $\underline{X}$: $\mathcal{B}$: Borel sets of $\mathbb{R}^p$.

$\nu_{\underline{X}} : \mathcal{B} \to [0, 1]$.

$\nu_{\underline{X}}(B) = Pr[\underline{X} \in B] \quad \forall B \in \mathcal{B}$.

Special cases: $\nu_{\underline{X}}(B) = \int_B f_{\underline{X}}(\underline{t})\, d\underline{t}$

$f_{\underline{X}}$: density of $\underline{X}$.

- Mean of $\underline{X}$: $E[\underline{X}] = \underline{\mu} := \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$

- Covariance of the components:

$$\sigma_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)], \quad \begin{matrix} j = 1 \dots p \\ k = 1 \dots p \end{matrix}$$

$\Sigma := [\sigma_{jk}]$, $p \times p$ matrix.

We can write Σ as:

$$\Sigma = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'].$$

We'll work with combinations (linear) of random vectors.

For this reason, it's comfortable to introduce the matrices:

$$V := \begin{bmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{pp} \end{bmatrix}, \quad \sigma_{ii} \geq 0$$

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\sigma_{pp}} \end{bmatrix} \qquad V^{-\frac{1}{2}} = \begin{bmatrix} 1/\sqrt{\sigma_{11}} & & 0 \\ & \ddots & \\ 0 & & 1/\sqrt{\sigma_{pp}} \end{bmatrix} \quad \text{if } \sigma_{ii} > 0.$$

$$\rho := V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

correlation matrix ($p \times p$ symmetric)

1) Consider now $\underline{c} \in \mathbb{R}^p$.

$\Rightarrow \underline{c}'\underline{X} = c_1 X_1 + c_2 X_2 + \dots + c_p X_p \in \mathbb{R}$.

- $E[\underline{c}'\underline{X}] = c_1 E[X_1] + \dots + c_p E[X_p] = \underline{c}'\underline{\mu}$.
- $\mathrm{Var}(\underline{c}'\underline{X}) = ?$

Exercise: $\mathrm{Var}(c_1 X_1 + c_2 X_2) = c_1^2 \mathrm{Var}(X_1) + c_2^2 \mathrm{Var}(X_2) + 2 c_1 c_2 \mathrm{Cov}(X_1, X_2)$.

$\Rightarrow$ I can reduce the uncertainty if $\mathrm{Cov}(X_1, X_2) \leq 0$.

In general, $\mathrm{Var}(\underline{c}'\underline{X}) = \underline{c}'\Sigma\underline{c}$.

2) If we want to take many linear combinations, consider:

C : $k \times p$ matrix of constants.

Consider $C\underline{X}$. $C = \begin{pmatrix} \underline{c}_1' \\ \vdots \\ \underline{c}_k' \end{pmatrix}$, $\underline{c}_i \in \mathbb{R}^p$. $\Rightarrow C\underline{X} = \begin{pmatrix} \underline{c}_1'\underline{X} \\ \vdots \\ \underline{c}_k'\underline{X} \end{pmatrix}$.

- $E[C\underline{X}] = C\underline{\mu}$.
- $\mathrm{Cov}(C\underline{X}) = C\Sigma C'$.

Particular case: $C = \underline{c}' = (c_1 \dots c_p)$ $1 \times p$.

$\mathrm{Cov}(C\underline{X}) = \mathrm{Var}(\underline{c}'\underline{X}) = C\Sigma C' = \underline{c}'\Sigma\underline{c}$.

Conclusion:

Data: $\mathbb{X}$ $n \times p$ $\xrightarrow{\text{inference}}$ Model: $\underline{X}$ random vector.

We use the data to estimate $\underline{\mu}$, Σ.

- Estimate for μ:

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \simeq \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} \leftarrow \text{generated by data}$$

But why does it work? We can see $\bar{\underline{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$ as a realization of the object $\frac{1}{n}\sum_{i=1}^{n} \underline{X}_i =: \bar{\underline{X}}$. (the estimator of $\underline{\mu}$).

Notation: { estimator: algorithm.
estimate: product of the algorithm.

Proposition: Let $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim \underline{X}$, $E[\underline{X}] = \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$. Then:

1. $E[\bar{\underline{X}}] = \underline{\mu}$ (i.e. $\bar{\underline{X}}$ is unbiased).
2. $\text{Cov}(\bar{\underline{X}}) = \frac{1}{n}\Sigma$.

Proof:

$$E[\bar{\underline{X}}] = E\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \underline{\mu}.$$

$$\text{Cov}(\bar{\underline{X}}) = E[(\bar{\underline{X}} - \underline{\mu})(\bar{\underline{X}} - \underline{\mu})'] = E\left[\left(\frac{1}{n}\Sigma \underline{X}_i - \underline{\mu}\right)\left(\frac{1}{n}\Sigma \underline{X}_i - \underline{\mu}\right)'\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[(\underline{X}_i - \underline{\mu})(\underline{X}_i - \underline{\mu})']$$

$$= E\left[\frac{1}{n}\Sigma(\underline{X}_i - \underline{\mu})\left(\frac{1}{n}\Sigma(\underline{X}_i - \underline{\mu})\right)'\right] =$$

$$= \frac{1}{n^2} E\left[\sum_{\substack{i=1 \\ k=1}}^{n} (\underline{X}_i - \underline{\mu})(\underline{X}_k - \underline{\mu})'\right] =$$

$$= \frac{1}{n^2}\sum_{i,k=1}^{n} E[(\underline{X}_i - \underline{\mu})(\underline{X}_k - \underline{\mu})'].$$

Now, $E[(\underline{X}_i - \underline{\mu})(\underline{X}_k - \underline{\mu})'] = \begin{cases} \Sigma, & i = k. \\ 0, & i \neq k \text{ (independence)} \end{cases}$

$$\Rightarrow \text{Cov}(\bar{\underline{X}}) = \frac{1}{n^2}\sum_{i=1}^{n}(\Sigma) = \frac{n}{n^2}\Sigma = \frac{1}{n}\Sigma. \quad \text{(end)}$$

- Estimate for Σ:

Exercise: Let $S := \frac{1}{n}\Sigma(\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})'$, estimator of Σ.

Prove that $E[S] = \frac{n-1}{n}\Sigma$, i.e. S is biased.

$\Rightarrow$ an unbiased estimator of Σ is $\frac{1}{n-1}S$.

17/3/20

Recap:

$$X = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} = [\underline{y}_1 \ \cdots \ \underline{y}_p] \qquad \underline{x}_i \in \mathbb{R}^p$$

data matrix

$\underline{x}_i$ is the realization of a random vector $\underline{X}_i$ in $\mathbb{R}^p$.

We will assume that $\underline{X}_1, \dots, \underline{X}_n$ are i.i.d $\sim \underline{X}$, with:

$$E[\underline{X}] = \underline{\mu}, \quad \text{cov}(\underline{X}) = \Sigma \quad \text{(unknown)}.$$

In order to estimate these parameters, we introduce:

$$\bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i \quad \text{random vector in } \mathbb{R}^p, \text{ called } \underline{\text{sample mean}}.$$

The realization of $\bar{\underline{X}}$ is $\bar{\underline{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$.

$$S' = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \quad \text{is a } p \times p \ \underline{\text{random matrix}}$$

The realization of S' is $S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ & & \vdots \\ & & s_{pp} \end{bmatrix}$, $\underline{\text{sample cov. matrix}}$.

What we've seen yesterday is that:

Proposition:

1) $E[\bar{\underline{X}}] = \underline{\mu}$, i.e. $\bar{\underline{X}}$ is $\underline{\text{unbiased}}$ for $\underline{\mu}$.

$\text{Cov}(\bar{\underline{X}}) = \frac{1}{n}\Sigma$, i.e. the covariance of $\bar{\underline{X}}$ is $\underline{\text{smaller than}}$ Σ.

2) $E[S'] = \frac{n-1}{n}\Sigma$, i.e. S' is a $\underline{\text{biased estimator}}$.

A simple corollary is: $E\left[\frac{n}{n-1} S'\right] = \Sigma$.

So $\frac{n}{n-1} S'$ is an $\underline{\text{unbiased estimator}}$ for Σ.

$$\frac{n}{n-1} S' = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})'(\underline{X}_i - \bar{\underline{X}})'.$$

From now on, we will denote:

$$\boxed{S' = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})'}.$$

When we'll need the "dd S'", i.e. the covariance of the sample, we will denote it with S_n.

Today we will focus on S'. Some alternative notations for S will be handy. Remind the definition of deviation vector $\underline{d}_j$:

$$\underline{d}_j = \underline{y}_j - \pi_{\underline{y}_j | \underline{1}} = \underline{y}_j - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}} \underline{y}_j = \underbrace{\left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)}_{\text{orthogonal projector on } \mathcal{L}^{\perp}(\underline{1})} \underline{y}_j$$

The deviation matrix is defined by:

$$d := [\underline{d}_1 \ \underline{d}_2 \ \cdots \ \underline{d}_p] \quad , \ n \times p \text{ matrix}.$$

We can see that:

$$d = \left(I - \frac{11'}{1'1} \right) \mathbb{X}.$$

Moreover,

$$S' = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ & & & \vdots \\ & & & s_{pp} \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} \sum (x_{i1} - \bar{x}_1)^2 & \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \cdots \\ \vdots \end{bmatrix} =$$

$$= \frac{1}{n-1} \begin{bmatrix} \underline{d}_1' \underline{d}_1 & \underline{d}_1' \underline{d}_2 & \cdots & \underline{d}_1' \underline{d}_p \\ & & & \vdots \\ & & & \underline{d}_p' \underline{d}_p \end{bmatrix} = \frac{1}{n-1} d'd =$$

$$= \frac{1}{n-1} \mathbb{X}' \left(I - \frac{11'}{1'1} \right)' \left(I - \frac{11'}{1'1} \right) \mathbb{X}.$$

Given that $I - \frac{11'}{1'1}$ is an orthogonal projection,

$$\left(I - \frac{11'}{1'1} \right)' \left(I - \frac{11'}{1'1} \right) = \left(I - \frac{11'}{1'1} \right).$$

$$\Rightarrow \quad \boxed{S = \frac{1}{n-1} \mathbb{X}' \left(I - \frac{11'}{1'1} \right) \mathbb{X}}.$$

Variability in a "multivariate" sense:

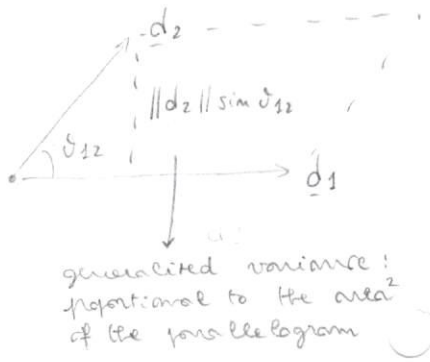
How do we capture the variability of multivariate data? We have two options:

1) Generalized variance: $\boxed{\text{Det}(S)}$.
2) Total Variance: $\boxed{\text{tr}(S)}$.

Example: consider $p = 2$. Then:

$$S = \frac{1}{n-1} d'd = \frac{1}{n-1} \begin{bmatrix} \underline{d}_1' \underline{d}_1 & \underline{d}_1' \underline{d}_2 \\ \underline{d}_2' \underline{d}_1 & \underline{d}_2' \underline{d}_2 \end{bmatrix} =$$

$$= \frac{1}{n-1} \begin{bmatrix} \|d_1\|^2 & \|d_1\| \|d_2\| \cos \vartheta_{12} \\ \|d_1\| \|d_2\| \cos \vartheta_{12} & \|d_2\|^2 \end{bmatrix}.$$



The generalized variance is:

$$\text{Det}(S') = \frac{\|d_1\|^2 \|d_2\|^2}{(n-1)^2} - \|d_1\|^2 \|d_2\|^2 \cos^2 \vartheta_{12} =$$

$$= \frac{\|d_1\|^2 \|d_2\|^2}{(n-1)^2} (1 - \cos^2 \vartheta_{12}) = \frac{\|d_1\|^2 \|d_2\|^2}{(n-1)^2} \sin^2 \vartheta_{12} =$$

$$\propto (\text{Area of parallelog } (\underline{d}_1, \underline{d}_2))^2.$$

Then, $\text{Det}(S) \uparrow \iff$ Area parallelogram $\uparrow$.

Moreover, $\text{Det}(S) = 0 \Rightarrow$ either $\begin{cases} \|d_1\| \text{ or } \|d_2\| = 0, \\ \vartheta_{12} = 0 \end{cases}$

The total variance, instead, is:

$$\text{tr}(S) = \left(\|d_1\|^2 + \|d_2\|^2 \right) \frac{1}{n-1}.$$

This holds for p in general:

$$\begin{cases} \text{Det}(S) \propto \text{Vol}^2(\text{parallelotype}(\underline{d}_1, \dots, \underline{d}_p)). \\ \text{trace}(S) \propto \|\underline{d}_1\|^2 + \|\underline{d}_2\|^2 + \dots + \|\underline{d}_p\|^2. \end{cases}$$

We need both these variances in order to describe the variability of the multivariate data.

Proposition:

$$\text{Det}(S) = 0 \iff \underline{d}_1, \dots \underline{d}_p \text{ are linearly dependent},$$

$$\text{i.e. } \exists \underline{c} \neq \underline{0} \text{ s.t. } d\underline{c} = \underline{0} \quad (\text{i.e. } c_1 \underline{d}_1 + \dots + c_p \underline{d}_p = \underline{0}).$$

Proof:

$\boxed{\Leftarrow}$ Suppose $\underline{d}_1, \dots, \underline{d}_p$ are linearly dependent, i.e.

$\exists \underline{c} \in \mathbb{R}^p$ s.t. $d\underline{c} = \underline{0}$.

Then: $S = \frac{1}{n-1} d'd \Rightarrow S\underline{c} = \frac{1}{n-1} d' \underbrace{d\underline{c}}_{=\underline{0}} = \underline{0}$

$\Rightarrow$ the columns of S are linearly dependent $\Rightarrow \text{Det}(S) = 0$.

$\boxed{\Rightarrow}$ Assume $\det S = 0$. Then $\exists \underline{c} \neq \underline{0}$ s.t. $S\underline{c} = \underline{0}$.

$$\Rightarrow \underline{c}' S \underline{c} = \underline{0} \Rightarrow \frac{1}{n-1} \underline{c}' d'd \underline{c} = 0, \text{ or } \frac{1}{n-1} \|d\underline{c}\|^2 = 0.$$

By property of norm, $\|d\underline{c}\| = 0 \Rightarrow d\underline{c} = \underline{0}$, i.e.

$\underline{d}_1, \dots, \underline{d}_p$ are linearly dependend. (end)

But what does it mean that $\underline{d}_1, \dots, \underline{d}_p$ are linearly dependent?

We have seen that $\exists \; c_1 d_1 + \dots + c_p \underline{d}_p = \underline{0}$, $\underline{c} \neq \underline{0}$.

W.l.o.g, $c_1 \neq 0$. Then:

$$\underline{d}_1 = - \sum_{i=2}^{p} \frac{c_i}{c_1} \underline{d}_i$$

$$\Rightarrow \boxed{\underline{y}_1 = \bar{x}_1 \cdot \underline{1} - \sum_{i=2}^{p} \frac{c_i}{c_1} (\underline{y}_i - \bar{x}_i \cdot \underline{1})}$$

Hence there is no new informations from variable X_1: we have a perfect linear relationship between the first variable and the other.

Proposition: Let $\mathbb{X}$ $n \times p$ data matrix.

If $p \geq n$, then $\det(S) = 0$.

Proof: $d = [\underline{d}_1, \dots, \underline{d}_p]$, $\underline{d}_i \in \mathbb{R}^n$, but $\underline{d}_i \in \mathcal{L}^1(\underline{1})$, $i = 1 \dots n$.

$\dim(\mathcal{L}^1(\underline{1})) = n-1$: $[\underline{d}_i]_{i=1}^p$ live in a space of dimension $n-1$. ($n-1$ degrees of freedom)

But $p \geq n \Rightarrow \underline{d}_1, \dots, \underline{d}_p$ are linearly dependent.

$\Rightarrow \det(S) = 0$.

(end)

Example: suppose our raw data is:

$$\begin{matrix} X & Y \end{matrix}$$
$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}.$$

Note that the data matrix $\mathbb{X}$ doesn't have to be equal to the raw data. For example:

$$\mathbb{X} = \begin{matrix} X & X^2 & \dots & X^p & Y \end{matrix}$$
$$\begin{bmatrix} x_1 & x_1^2 & \dots & x_1^p & y_1 \\ x_2 & x_2^2 & & x_2^p & y_2 \\ \vdots & & & & \vdots \\ x_n & x_n^2 & \dots & x_n^p & y_n \end{bmatrix}.$$

Taking $p = n-1$, then the number of columns of $\mathbb{X}$ is $p+1 = n$.

$\Rightarrow \det(S) = 0$. Assume that $x_n, x_n^2, \dots x_n^p$ are linearly independent.

Then: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p$.

↳ we have a perfect fit $\Rightarrow$ overfitting! Remember the bias-variance tradeoff.

This example tells us we'll always be able to find a perfect fit for the data, but that's not what we want.

Spectral Decomposition of S:

We have seen that a necessary condition for $\det(S) \neq 0$ is $n \geq p+1$.

We know that S is a $p \times p$ symmetric matrix, and $s_{ij} \in \mathbb{R}$ $\forall i, j = 1, \dots, p$.

From linear algebra, we do know that:

$$\exists\ \lambda_1, \dots, \lambda_p \in \mathbb{R},$$
$$\underline{e}_1, \dots, \underline{e}_p \in \mathbb{R}^p \quad \text{s.t.} \quad \underline{e}_i' \underline{e}_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

$$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$$

spectral decomposition of S

In fact, $(\lambda_i, \underline{e}_i)$ is the eigenvalue - eigenvector couple of S.

This decomposition is the core of dimensional reduction.

Let: $P := [\underline{e}_1, \dots, \underline{e}_p]$, $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$.

Since $S \underline{e}_i = \lambda_i \underline{e}_i \quad \forall i = 1, \dots, p$, then:

$$\boxed{S = P \Lambda P'}$$ spectral decomposition of S (matrix form)

moreover,

$$\begin{cases} \det(S) = \prod_{i=1}^{p} \lambda_i & \text{generalized variance} \\ \operatorname{tr}(S) = \sum_{i=1}^{p} \lambda_i & \text{total variance} \end{cases}$$

Proposition: S is positive semi-definite.

If $\det(S) \neq 0$, then S is positive definite.

Proof: need to prove that $\underline{c}' S \underline{c} \geq 0 \quad \forall \underline{c} \in \mathbb{R}^p$.

But $\underline{c}' S \underline{c} = \frac{1}{n-1} \underline{c}' D' D \underline{c} = \frac{1}{n-1} \|D\underline{c}\|^2 \geq 0$.

Suppose that $\exists \underline{c} \neq \underline{0}$ s.t. $\underline{c}' S \underline{c} = 0$. $\Rightarrow \|D\underline{c}\| = 0$.

$\Rightarrow D\underline{c} = \underline{0} \Rightarrow \underline{d}_1 \dots \underline{d}_p$ are lin. dep. (prop. before) $\Rightarrow \det(S) = 0$.

So if $\det(S) \neq 0$, then S is positive definite ($\det(S) > 0$). (end)

Mahalanobis Distance and its geometry:

Notation: from now on, writing $S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$, we assume that:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \quad (\geq 0)$$ ordering eigenvalues of S

λ_1 | λ_2 | λ_p

$\underline{e}_1$ $\underline{e}_2$ $\underline{e}_p$

Assume that $\det(S) \neq 0 \Rightarrow \lambda_1 \geq \dots \geq \lambda_p > 0$.

So we have:

$$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$$

$$\Rightarrow \boxed{S^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'}.$$

Then, S^{-1} induces a metric on $\mathbb{R}^p$. Indeed, let $d_{S^{-1}}$ be defined as:

$$\forall \underline{x}, \underline{y} \in \mathbb{R}^p: \quad \boxed{d^2_{S^{-1}}(\underline{x}, \underline{y}) := (\underline{x} - \underline{y})' S^{-1} (\underline{x} - \underline{y})}$$ Mahalanobis' distance

$d_{S^{-1}}$ is a distance in $\mathbb{R}^p$ [prove it!].

Consider now the set:

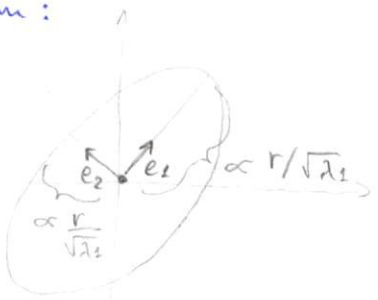
$$\mathcal{E}_{r^2, S^{-1}}(\bar{x}) := \left\{ x \in \mathbb{R}^p : d^2_{S^{-1}}(x, \bar{x}) \leq r^2 \right\}$$

$$= \left\{ x \in \mathbb{R}^p : (x - \bar{x})' S^{-1} (x - \bar{x}) \leq r^2 \right\}$$

What are the characteristics of this set?

Consider, in general, B positive definite of the form:

$$B = \sum_{i=1}^{p} \lambda_i e_i e_i'$$



Consider the points $x' B x = r^2$.

$\Rightarrow$ These points are an ellipse.

In our case, $\mathcal{E}_{r^2, S^{-1}}(x)$ is the following ellipse:

The axes of $\mathcal{E}_{S^{-1}}(x, r^2)$ are proportional to the squared root of the eigenvalues of S.

The volume of $\mathcal{E}_{r^2, S^{-1}}(x)$ is:

$$Vol\left(\mathcal{E}_{r^2, S^{-1}}(x)\right) = K_p r^p \sqrt{\prod_{i=1}^{p} \lambda_i} = K_p r^p \sqrt{Det(S)}$$

generalized variance

Example: let $p = 2$, $S = \begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix}$, $s_{11} > s_{22}$.

Take q_1, q_2 s.t.: $q_1 = \begin{pmatrix} q_1 \\ \bar{x}_2 \end{pmatrix}$, $q_2 = \begin{pmatrix} \bar{x}_1 \\ q_2 \end{pmatrix}$,

and $d_{euc}(\bar{x}, q_1) = d_{euc}(\bar{x}, q_2)$.

Anyway, q_1 is a "normal point" with respect to the cloud, while q_2 is an extraordinary point.

We want to deal with standardized data:

$$\begin{cases} d_{euc}(std(q_1), std(\bar{x})) = \dfrac{|q_1 - \bar{x}_1|}{\sqrt{s_{11}}} \\ d_{euc}(std(q_2), std(\bar{x})) = \dfrac{|q_2 - \bar{x}_2|}{\sqrt{s_{22}}} \end{cases}$$

Given that $|q_1 - \bar{x}_1| = |q_2 - \bar{x}_2|$ and $s_{11} > s_{22}$, then:

$$d_{euc}(std(q_1), std(\bar{x})) < d_{euc}(std(q_2), std(\bar{x}))$$

In general, taking $q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$, we have:

$$d_{euc}(std(q), std(\bar{x})) = \sqrt{\frac{(q_1 - \bar{x}_1)^2}{s_{11}} + \frac{(q_2 - \bar{x}_2)^2}{s_{22}}} = \sqrt{(q - \bar{x})^T S^{-1} (q - \bar{x})} =$$

$$= d_{S^{-1}}(q, \bar{x}).$$

Taking a more general shape of $S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$, we have an inclined ellipse.

19/3/20

Recap:

- X : data matrix $n \times p$,
- $\bar{x}$: sample mean $\in \mathbb{R}^p$,
- S : sample covariance $p \times p$.

If $\det(S) > 0$, $d_{S^{-1}} : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty]$.

$$d^2_{S^{-1}} := (x - y)' S^{-1} (x - y)$$

Mahalanobis' distance

Example: ($p = 2$) $S = \begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix}$.

$$\mathcal{E}_r(\bar{x}) = \left\{ x \in \mathbb{R}^p : (x - \bar{x})' S^{-1} (x - \bar{x}) \le r^2 \right\}.$$

↳ ellipse of "radius" r centered in $\bar{x}$.

$$\text{Area}(\mathcal{E}_r(\bar{x})) \propto r^2 \sqrt{\text{Det}(S)} = r^2 \sqrt{s_{11} \cdot s_{22}}.$$

The larger the variability $\text{Det}(S)$, the larger this neighborhood.

If $\text{Det}(S)$ or $\text{tr}(S)$ ↑ $\Rightarrow$ $\text{Area}(\mathcal{E}_r(\bar{x}))$ ↑ (with r fixed)

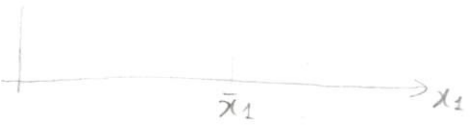
One could say that it's a special case. Is this really special?

- it's easy to extend to $p > 2$ (use your imagination).
- what if S is not diagonal? Let $p \ge 1$, any S : $\text{Det}(S) > 0$.

(In the figure, we have positive correlation between X_1, X_2)

Consider the spectral decomposition of S:

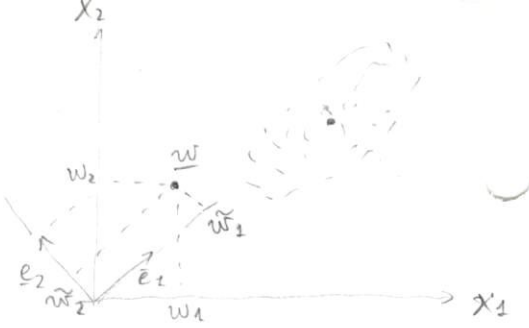
$$S = \sum_{i=1}^{p} \lambda_i e_i e_i' \quad , \quad (\lambda_1 \ge \lambda_2 \ge \dots \lambda_p > 0).$$



Or, in other words:

$$S = P \Lambda P' .$$

We'll introduce a new orthonormal system made by the eigenvectors of S.



Old system:

$$\underline{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix}$$

New system:

$$\underline{\tilde{w}} = \begin{pmatrix} \underline{e}_1' \underline{w} \\ \vdots \\ \underline{e}_p' \underline{w} \end{pmatrix} = P' \underline{w} .$$

Since the length doesn't change:

$$\|\underline{w}\|^2 = \underline{w}' \underline{w} = \underline{\tilde{w}}' \underline{\tilde{w}} = \|\underline{\tilde{w}}\|^2 = \underline{w}' P' P \underline{w} .$$

We can do the same with our data:

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \qquad \tilde{\mathbb{X}} = \begin{bmatrix} (P' \underline{x}_1)' \\ \vdots \\ (P' \underline{x}_n)' \end{bmatrix} = \begin{bmatrix} \underline{x}_1' P \\ \vdots \\ \underline{x}_n' P \end{bmatrix} = \mathbb{X} P .$$

Let's analyze the covariance matrix:

$$S = \frac{1}{n-1} \mathbb{X}' \left(I - \frac{1 1'}{1' 1} \right) \mathbb{X} ,$$

$$\tilde{S} = \frac{1}{n-1} \tilde{\mathbb{X}}' \left(I - \frac{1 1'}{1' 1} \right) \tilde{\mathbb{X}}$$

$$= \frac{1}{n-1} P' \mathbb{X}' \left(I - \frac{1 1'}{1' 1} \right) \mathbb{X} P =$$

$$= P' S P = P' P \Lambda P' P =$$

$$= \Lambda .$$

$$\Rightarrow \tilde{S} = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} .$$

In this new system, the covariances are zero!

Take-home message: there is always a reference system for which the coordinates are uncorrelated.

"We need to look at the data with particular glasses".

Let's now analyze the generalized variance:

$$\mathrm{Det}(\tilde{S}) = \mathrm{Det}(\Lambda) = \prod_{i=1}^{p} \lambda_i = \mathrm{Det}(S) .$$

Same thing goes for total variance:

$$\mathrm{tr}(\tilde{S}) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i = \mathrm{tr}(S) .$$

$\Rightarrow$ The total variability doesn't depend on the reference system.

Remark:

$$P'P = \begin{pmatrix} e_1' \\ \vdots \\ e_p' \end{pmatrix} (e_1 \dots e_p) = \begin{pmatrix} e_1'e_1 & e_1'e_2 & \dots & e_1'e_p \\ & 1 & & 0 \\ & & 1 & \\ & 0 & & 1 \end{pmatrix} = I.$$

Example: (powerpoint). Correlation of the Milky Way.

↳ correlation is in the eyes of the beholder, not in the system itself!

We got a projection of a 1D linear space: it is the first principal component!

Principal Component Analysis (PCA):

Let X random vector in $\mathbb{R}^p$, $E[X] = \mu$, $\text{cov}[X] = \Sigma$.

Take $a \in \mathbb{R}^p$: $a'X = a_1 X_1 + \dots + a_p X_p$.

Problem: find a s.t. $\text{Var}(a'X)$ is maximum.

Is the problem well-posed?

Suppose that the solution is $a' = (10, 0, \dots, 0)'$, s.t. max. var $= a'X = 10 X_1$.

In this way, we could take:

$10 \cdot a = (100, 0, \dots, 0)$.

$\Rightarrow \text{Var}(10\, a'X) = 100\, \text{Var}(a'X)$.

So a cannot be the maximum!

Therefore, the problem is not well-posed. Let's reform it a little bit.

Problem: find a s.t. $\|a\| = 1$ and $\text{Var}(a'X)$ is max.

Formally:

$$\max_{a \in \mathbb{R}^p : \|a\| = 1} \text{Var}(a'X). \qquad (*)$$

$\text{Var}(a'X) = a'\Sigma a$, so that $(*)$ becomes:

$$\max_{a \in \mathbb{R}^p : \|a\| = 1} a'\Sigma a = \max_{a \in \mathbb{R}^p} \frac{a'\Sigma a}{a'a}, \quad \text{so} \quad \left\| \frac{a}{\|a\|} \right\| = 1.$$

Lemma: Let B $p \times p$ positive semi-definite matrix, and $B = \sum_{i=1}^{p} \lambda_i e_i e_i'$ its spectral decomposition. Then:

1) $\max_{x \in \mathbb{R}^p} \frac{x'Bx}{x'x} = \lambda_1$, $\quad \arg\max_{x \in \mathbb{R}^p} \frac{x'Bx}{x'x} = e_1$.

2) $\max_{x \in \mathbb{R}^p : x \perp e_1} \frac{x'Bx}{x'x} = \lambda_2$, $\quad \arg\max \dots = e_2$.

p) $\max_{\underline{x} \in \mathbb{R}^p :} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \lambda_p = \min_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}}$.

$\underline{x} \perp \underline{e}_1, \dots, \underline{e}_{p-1}$

Proof: $B = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i' = P \Lambda P'$.

1) $\frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \frac{\underline{x}' P \Lambda P' \underline{x}}{\underline{x}' P P' \underline{x}} = \left[\text{call } \underline{y} = P' \underline{x} \right]$

$= \frac{\underline{y}' \Lambda \underline{y}}{\underline{y}' \underline{y}} = \frac{\sum_{i=1}^{p} \lambda_i y_i^2}{\sum_{i=1}^{p} y_i^2} \leq \lambda_1 \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} y_i^2} = \lambda_1. \Rightarrow \lambda_1 = \max \dots$

Take now $\underline{x} = \underline{e}_1$. Then:

$\frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \frac{\underline{e}_1' B \underline{e}_1}{\underline{e}_1' \underline{e}_1} = \lambda_1 \underline{e}_1' \underline{e}_1 = \lambda_1 \Rightarrow \underline{e}_1 = \arg\max \dots$

2) $\max_{\underline{x} \in \mathbb{R}^p : \underline{x} \perp \underline{e}_1} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \left\{ \underline{y} = P' \underline{x} = \begin{pmatrix} \underline{e}_1' \\ \underline{e}_2' \\ \vdots \\ \underline{e}_p' \end{pmatrix} \underline{x} \overset{\underline{x} \perp \underline{e}_1}{=} \begin{pmatrix} 0 \\ \underline{e}_2' \underline{x} \\ \vdots \\ \underline{e}_p' \underline{x} \end{pmatrix} \right\} =$

$= \frac{\underline{y}' \Lambda \underline{y}}{\underline{y}' \underline{y}} = \frac{\sum_{i=2}^{p} \lambda_i y_i^2}{\sum_{i=2}^{p} y_i^2} \leq \lambda_2. \Rightarrow \max \dots = \lambda_2$

Taking $\underline{x} = \underline{e}_2$, $\frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \frac{\underline{e}_2' B \underline{e}_2}{\underline{e}_2' \underline{e}_2} = \lambda_2 \underline{e}_2' \underline{e}_2 = \lambda_2$.

$\Rightarrow \arg\max \dots = \underline{e}_2$.

p) Same for max, argmax.
In order to prove the minimum:

$\frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \frac{\underline{y}' \Lambda \underline{y}}{\underline{y}' \underline{y}} = \frac{\sum_{i=1}^{p} \lambda_i y_i^2}{\sum y_i} \geq \lambda_p \Rightarrow \lambda_p = \min \dots$

$\underline{x} = \underline{e}_p \Rightarrow \underline{e}_p = \arg\min \dots$ (end)

Using the lemma, if $\Sigma = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$, then:

$\max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \operatorname{Var}(\underline{a}' \underline{X}) = \lambda_1.$

$\arg\max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \operatorname{Var}(\underline{a}' \underline{X}) = \underline{e}_1.$

Definition: $y_1 := \underline{e}_1' \underline{X}$ is called first principal component (PC1).

(Note: $y_1 = \underline{e}_1' (\underline{X} - \underline{\mu})$.)

Problem: $\max_{a \in \mathbb{R}^p : \|a\|=1,\ \mathrm{Cov}(a'X, e_1'X)=0} \mathrm{Var}(a'X) = \max_{a \in \mathbb{R}^p : \mathrm{Cov}(a'X, e_1'X)=0} \frac{a'\Sigma a}{a'a}$. (*)

How can we express $\mathrm{Cov}(a'X, b'X)$?

Taking C $k \times p$, $\mathrm{Cov}(CX) = C \Sigma C'$.

If $C = \begin{bmatrix} a' \\ b' \end{bmatrix}$ $2 \times p$, then: $\mathrm{Cov}(CX) = \mathrm{Cov}\left(\begin{pmatrix} a'X \\ b'X \end{pmatrix}\right) = \begin{bmatrix} a' \\ b' \end{bmatrix} \Sigma \begin{bmatrix} a & b \end{bmatrix}$

$$= \begin{bmatrix} a'\Sigma a & a'\Sigma b \\ b'\Sigma a & b'\Sigma b \end{bmatrix}.$$

$\Rightarrow \mathrm{Cov}(a'X, b'X) = a'\Sigma b$. Hence:

$\mathrm{Cov}(a'X, e_1'X) = a'\Sigma e_1 = \lambda_1 a' e_1 = 0 \iff \boxed{a \perp e_1}$

Then we can reformulate (*) as:

$$\begin{cases} \max_{a \in \mathbb{R}^p : a \perp e_1} \frac{a'\Sigma a}{a'a} = \lambda_2 \quad \text{(Lemma)} \\ \operatorname{argmax}_{\dots} \dots = e_2. \end{cases}$$

Definition: $\boxed{y_2 = e_2'X}$ is PC2 (a second principal component).

(Note: other possibility is to take $y_2 = e_2'(X - \mu)$.

We can generalize easily the solution of the problem for $j \geq 2$:

General problem:

$$\max_{\substack{a \in \mathbb{R}^p : \\ \|a\|=1, \\ \mathrm{Cov}(a'X, e_i'X)=0 \\ \forall i = 1, \dots, j-1}} \mathrm{Var}(a'X) = \lambda_j \quad , \quad \operatorname{argmax}_{\dots} \dots = e_j.$$

Definition: $\boxed{Y_j = e_j'X}$ is PC$_j$ (j-th principal component).

($Y_j = e_j'(X - \mu)$)

$\boxed{Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = P'X}$ is the vector of p-principal components.

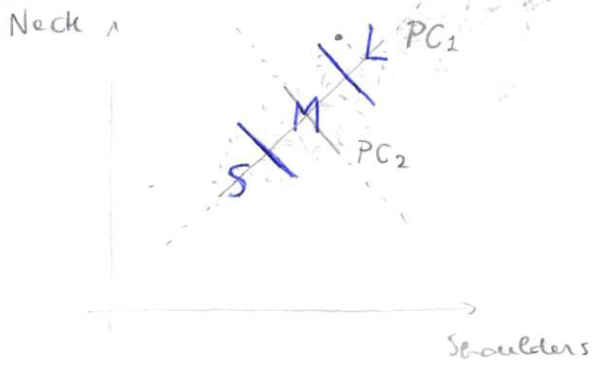
Proposition:

- $E[Y] = E[P'X] = P'\mu$. [if $Y = P'(X - \mu) \Rightarrow E[Y] = 0$.]
- $\mathrm{Cov}(Y) = \mathrm{Cov}(P'X) = P'\Sigma P = P'P\Lambda P'P = \Lambda$.

So: $\begin{cases} \text{cov}(Y_i, Y_j) = 0 & \text{if } i \neq j, \; i,j = 1 \dots p. \\ \text{Var}(Y_i) = \lambda_i. \end{cases}$

In this way, I've constructed a powerful tool of dimension reduction.

Example: you are a tailor, and you measure size of neck & shoulders.



We can explain most of variability only with PC_1. So we project all the clients along PC_1, and divide PC_1 in three regions, $\{S, M, L\}$.
You want a better shirt? Go to a tayler! :)

Observations:

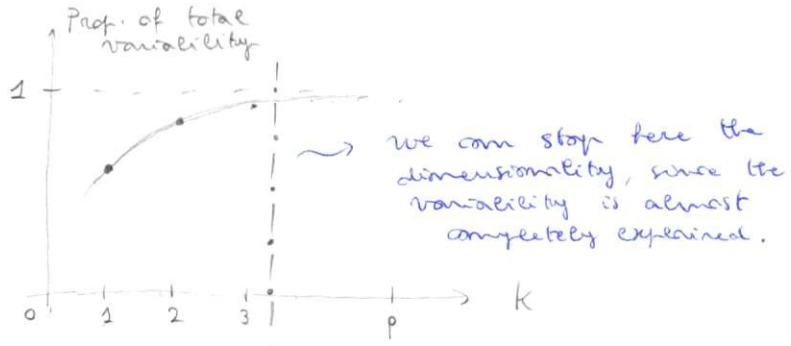
1) $\text{General. Variance}(\underline{Y}) = \text{Det}(\Lambda) = \prod_{i=1}^{p} \lambda_i = \text{Det}(\Sigma) = \text{G.V.}(\underline{X})$.
($\text{G.V.}(\underline{Y})$)

$\text{Total Variance}(\underline{Y}) = \text{tr}(\Lambda) = \text{tr}(\Sigma) = \text{total variance}(\underline{X})$.

We are not losing anything in the overall variance of the system.

2) $\text{Var}(Y_1) = \lambda_1$: Y_1 is capturing $\frac{\lambda_1}{\sum_i \lambda_i}$ proportion of total variability.

$\text{Var}(Y_2) = \lambda_2$: $\{Y_1, Y_2\}$ are capturing $\frac{\lambda_1 + \lambda_2}{\sum_i \lambda_i}$ of total variability.



Remark: interpretation of the PCs.
Going back to the tailor, what is the score $\{S, M, L\}$ speaking about? Neck or shoulders? It's a combination of the two: it doesn't have to make sense physically, it explains the variability of data.

Lab session 2. (finished Lab. 1). 20/3/20

p-value: suppose that $\beta_1 = 0$, so H_0 hypothesis is true.
Then what is the probability of observing an estimate such $\hat{\beta}_1$?
That probability is the p-value.
If p-value is small, then what we saw is "a miracle"
if H_0 is true: H_0 is not acceptable.

Take-home message: exploring the data with graphic tools is very useful.

Aneurysm: functional Principal Components.

Book: "How to lie with Statistics".

PCA (Continued): 23/3/20

$\underline{X}$: random vector, $E[\underline{X}] = \mu$, $Cov(\underline{X}) = \Sigma$.

$$\Sigma = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i' \quad \text{spectral decomposition}$$

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$.

$Y_i = \underline{e}_i' \underline{X}$: i-th principal component.

$Y_i = e_{1i} X_1 + \dots + e_{pi} X_p$.

($e_{1i}, \dots, e_{pi}$: loadings)

Proposition: $$Cov(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1 \dots p.$$

Proof:

$$Cor(Y_i, X_k) = \frac{Cov(Y_i, X_k)}{\sqrt{\lambda_i \cdot \sigma_{kk}}}.$$

$$Cov(Y_i, X_k) = Cov(\underline{e}_i' \underline{X}, \underline{u}_k' \underline{X}), \quad \underline{u}_k := (0 \dots 0 \underset{k}{1} 0 \dots 0).$$

$$\overset{!}{=} \underline{e}_i' \Sigma \underline{u}_k = \underline{u}_k' \Sigma \underline{e}_i =$$

$$\overset{!}{=} \lambda_i \underline{u}_k' \underline{e}_i = \lambda_i e_{ki}. \Rightarrow Cov(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}.$$

e_{ki}: k-th component of the i-th eigenvector.

(end)

Principal components obtained from standardized values:

$$\underline{Z} = V^{-\frac{1}{2}} (\underline{X} - \underline{\mu}) = \left(\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \dots, \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}} \right)'.$$

Standardized Random Vector

$$\Rightarrow \quad E[\underline{Z}] = \underline{0}, \quad \text{cov}(\underline{Z}) = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = \rho.$$

In this way, I can get rid of the unit of measures.

We can express $\rho = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$, where now $\underline{e}_i$ are eigenvectors of ρ and λ_i are eigenvalues of ρ, correlation matrix of $\underline{X}$.

The analysis proceeds as before:

$$Y_i = \underline{e}_i' \underline{Z} = \underline{e}_i' V^{-\frac{1}{2}} (\underline{X} - \underline{\mu}).$$ i-th principal component of $\underline{Z}$

We notice that:

1) $\sum_{i=1}^{p} \text{Var}(Y_i) = \text{tr}(\rho) = p = \sum_{i=1}^{p} \text{Var}(Z_i)$.

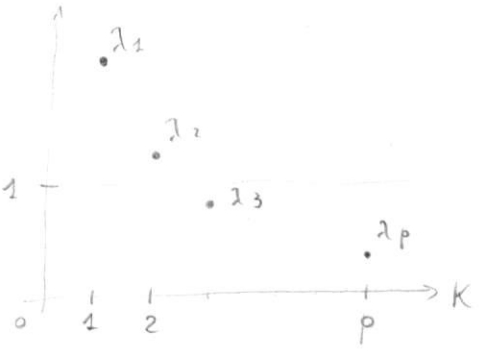
2) $\text{Cor}(Y_i, Z_k) = \frac{e_{ki} \sqrt{\lambda_i}}{1} = e_{ki} \sqrt{\lambda_i}$.

3) Proportion of total variability explained by the first k components:

$$\sum_{i=1}^{k} \lambda_i / p \qquad k = 1, \dots, p.$$

4) $\bar{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p} = \frac{\text{tr}(\rho)}{p} = \frac{p}{p} = 1$.

$\Rightarrow$ The average value of the eigenvalues in ρ is 1.
This motivates us to choose Y_i if $\lambda_i > 1$.



Example: $\Sigma = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$.

$$\Rightarrow \rho = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

$\Sigma_{22} >> \Sigma_{11}$. So doing the PCA in Σ could not be the best thing.
For example:

X_1 in cm.
X_2 in mm.

In this situation, the variability of the higher range will totally mask the variability of the other variable.